Case-Based Characterization and Analysis of Subgroup Patterns

Martin Atzmueller
University of Würzburg,
Department of Computer Science,
Am Hubland, 97074 Würzburg, Germany
atzmueller@informatik.uni-wuerzburg.de

Abstract

In this paper, we propose a case-based approach for characterizing and analyzing subgroup patterns: We present techniques for retrieving characteristic factors and cases, and merge these into prototypical cases for presentation to the user.

In general, cases capture knowledge and concrete experiences of specific situations. By exploiting case-based knowledge for characterizing a subgroup pattern, we can provide additional information about the subgroup extension. We can then present the subgroup pattern in an alternative condensed form that characterizes the subgroup, and enables a convenient retrieval of interesting associated (meta-)information.

1 Introduction

Subgroup discovery is a powerful and broadly applicable technique aiming at discovering interesting subgroups concerning a certain target property of interest, e.g., in the subgroup of smokers with a positive family history the risk of coronary heart disease (target property) is significantly higher than in the general population. The discovered interesting subgroups denote *nuggets* or *chunks* of knowledge. A subgroup is usually easy to interpret depending on a suitable description language, e.g., using conjunctive selection expressions. In that sense the subgroup description defining the subgroup objects (cases) stands for itself. Nevertheless, methods for subgroup characterization and analysis can be very useful, e.g., [Gamberger *et al.*, 2005; 2003], since they can be used to obtain further information about the extension of the subgroup, i.e., the cases covered by the subgroup description.

In the context of experience management [Bergmann, 2002] and case-based reasoning, cases contain specific knowledge of previously experienced, concrete problem situations [Aamodt and Plaza, 1994]. Usually, a case consists of a problem description part, a solution part, and additional attached meta-information, e.g., a description of the context of the case [Bartsch-Spörl *et al.*, 1999]. For example, in the medical domain specific cases for patients are collected which do not only include the case description (given by a set of attribute values) but also additional information, e.g., images from x-ray or sonographic examinations. Then, presenting a characteristic set of cases can be used for identifying typical problem situations and contexts of a specific subgroup. Such introspective information can support the user in interpreting the discovered subgroup patterns, by presenting a subgroup in an alternative form.

In this context, we propose case-based methods providing characterization and analysis capabilities concerning the subgroup extension, i.e., the cases covered by the subgroup. First, characteristic factors of the subgroup and their respective strengths are identified. Then, typical and extreme cases characterizing the subgroup are retrieved. The obtained set of factors, the respective cases, and associated meta-information contained in the cases, can then be provided as important additional information. For example, in the medical domain meta-information such as medical images, the name of the examiner that examined or documented a case, and a typical context of a subgroup pattern can both provide important analytical information and increase the actionability of the pattern.

In this paper we show how to characterize a subgroup in terms of its characteristic factors, how we can locate relevant characteristic cases using that information, and how we can finally summarize these by generating a *prototypical pattern case* containing the characteristic cases and factors. This case is then presented to the user as a representative case for a specific subgroup pattern.

The rest of the paper is organized as follows: We first introduce subgroup discovery, subgroup patterns, and characterization techniques in Section 2. After that, we present methods for case-based characterization and analysis of subgroup patterns in Section 3: We discuss an approach for obtaining a ranked list of the characteristic factors of a subgroup pattern. Next, we show how to analyze and exemplify subgroup patterns using typical and extreme cases. Based on these techniques, we present a method for generating *prototypical pattern cases* as a condensed representation of the factors and cases characterizing a given subgroup pattern. Next, we provide two case-studies in the medical domain in Section 4. Finally, we conclude the paper with a summary in Section 5, and point out interesting directions for future work.

2 Subgroup Discovery and Subgroup Patterns

The main application areas of subgroup discovery [Klösgen, 1996; Wrobel, 1997] are exploration and descriptive induction, to obtain an overview of the relations between a (dependent) target variable and a set of (independent) explaining variables. A subgroup pattern is specified by a subgroup description language; its quality is determined by a suitable quality function and a specific target variable (concept of interest).

In the following we first introduce the used knowledge representation, before we introduce subgroup patterns and describe a method for their statistical characterization.

2.1 General Definitions

First, let us introduce some vocabulary for the used knowledge representation: Let Ω_A be the set of all attributes. For each attribute $a \in \Omega_A$ a range $dom(a)$ of values is defined; we assume $\mathcal{V}_A$ to be the (universal) set of attribute values of the form $(a = v)$, $a \in \Omega_A, v \in dom(a)$. Other common names for attribute values are *findings* and *observations*.

A *case* c is defined as a tuple

$$c = (\mathcal{V}_c, \mathcal{I}_c),$$

where $\mathcal{V}_c \subseteq \mathcal{V}_A$ is the set of attribute values observed in the case c. The set of attribute values is also often called the set of *observations* for the given case, but can also include the solution of a case, e.g., a diagnosis in the medical domain. The set $\mathcal{I}_c$ provides additional (meta-) information.

In our context, we do not explicitly consider the solution part of a case that is usually modeled for case-based reasoning applications. It is easy to see, that the solution of a case, e.g., a diagnosis in medical domains, could easily be included in either the set $\mathcal{V}_c$ or the set $\mathcal{I}_c$, depending on the requirements of the application.

The set of all possible cases for a given problem domain is denoted by Ω_C. Let $CB \subseteq \Omega_C$ be the case base containing all available cases (also often called instances).

2.2 Subgroup Patterns

A subgroup pattern is defined by a subgroup description language. A single-relational propositional subgroup description

$$sd = \{e_1, e_2, \ldots, e_n\},$$

is defined by the conjunction of a set of selection expressions (selectors) $e_i = (a_i, V_i)$, i.e., selections on domains of attributes, $a_i \in \Omega_A, V_i \subseteq dom(a_i)$. For example, the subgroup given in the introduction is defined by the selectors *smoker=yes* and *family history=positive* (with respect to the target property *coronary heart disease*). The selection expressions contained in the subgroup description are also called the *principal factors* of the subgroup. We define Ω_E as the set of all selection expressions and Ω_{sd} as the set of all possible subgroup descriptions

The interestingness of a subgroup pattern can be flexibly formalized by a (user-defined) quality function

$$q : \Omega_{sd} \rightarrow R,$$

e.g., [Klösgen, 1996], that is used in order to evaluate a subgroup description $sd \in \Omega_{sd}$. Typical criteria for ranking subgroups and for estimating their quality include the difference in the distribution of the target variable concerning the subgroup and the general population, and the subgroup size. Usually the k best subgroups and/or the subgroups with a quality above a minimum threshold are selected.

2.3 Statistical Characterization of Subgroup Patterns

Subgroups can always be characterized by the factors used to describe them, i.e., by the selectors contained in the subgroup description. However, besides these *principal factors* there are certain *supporting factors* that can also be applied in order to characterize a subgroup, c.f., [Gamberger *et al.*, 2005]: The supporting factors are given by attribute values $supp \subseteq \mathcal{V}_A$ contained in the subgroup that are identified using basic statistical analysis. The value distributions of their corresponding (supporting) attributes differ significantly comparing the subgroup and the total population with respect to the concept of interest.

Thus, given a binary target variable, a supporting attribute a of a subgroup s is defined as an attribute with a significantly different distribution comparing the true positive (target class) cases contained in the subgroup s and all the negative (non-target) cases contained in the total population.

We say, that an attribute value $(a = v)$ corresponding to the selector $e = (a, \{v\})$ of a supporting attribute a is characteristic for the subgroup, i.e., it is a supporting factor, if it is positively associated with the true positive (target class) cases contained in the subgroup compared to all the negative cases. For testing the statistical significance of an attribute and an attribute value we apply the standard χ^2-test for independence with a 0.05 significance level (i.e., with a confidence level of 95%), and the correlation- or ϕ-coefficient for binary variables, respectively.

The principal factors can be regarded as *strong* factors, while the supporting factors can be regarded as a kind of *weak* factors: The principal factors are observed in all cases of a subgroup while the supporting factors are only observed in some cases. Nevertheless, the supporting factors can provide important additional information with respect to the target cases contained in the subgroup. As discussed by Gamberger et al. [Gamberger *et al.*, 2005] presenting the supporting factors characterizing the subgroup in addition to the principal factors can be very helpful for the user: Given the principal factors the supporting factors can provide additional evidence with respect to the target concept. In this way, observing the supporting factors can facilitate an easier recognition of target cases [Lavrac *et al.*, 2002]: If a case is assigned to a subgroup based on the principal factors, then observing a supporting factor provides for some evidence that the case is potentially positive with respect to the concept of interest. Thus, the supporting factors are used to point at specific characteristics of the target space covered by the subgroup. Then, we can define a generalized set F of *characteristic factors* as the union of the principal and supporting factors.

Considering the subpopulation defined by the subgroup the principal factors are contained in all cases. The supporting factors do not occur in all cases of the subgroup but may occur in many cases. Then, their individual strength in confirming the concept of interest, i.e., their *relative importance* can be scored. We will describe such an approach in Section 3.1 below.

3 Case-Based Subgroup Characterization and Analysis

In this section we describe the methods of the proposed approach for case-based subgroup characterization and analysis: Given a specific subgroup pattern, we first obtain a set of characteristic factors (selectors) for the subgroup. Next, we rank these factors and obtain a set of exemplifying cases for the given factors. After that, we create a *prototypical pattern case* capturing the characteristic factors of the subgroup pattern, the set of characteristic and exemplifying cases, and a set of relevant additional factors. The generated prototypical pattern case contains a set of (real) cases associated with the set of factors characterizing the subgroup and a selection of relevant additional factors contained in the set of cases, besides the characteristic factors. In that sense, the prototypical pattern case provides a representative summary of the characteristic factors and the respective retrieved cases for a specific subgroup pattern.

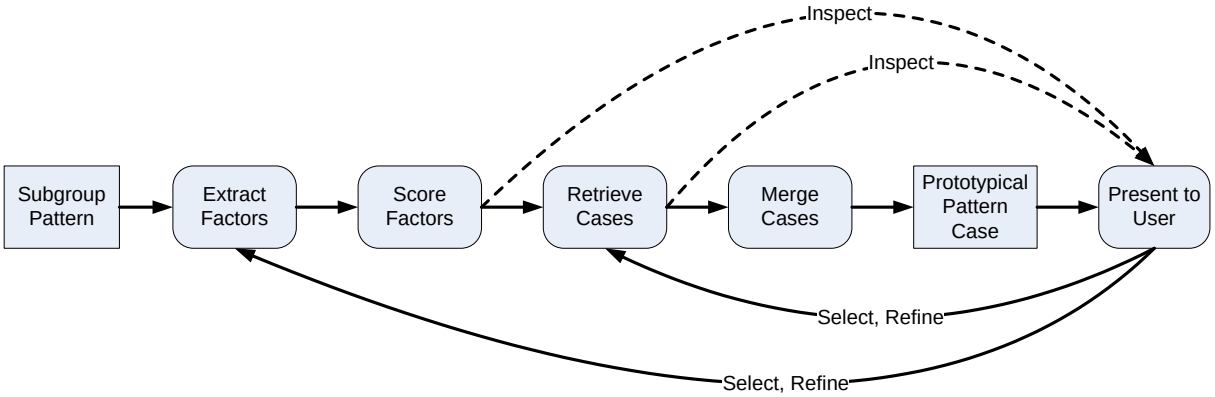

Figure 1: Process model: Case-based characterization and analysis

The approach for case-based characterization and analysis of subgroup patterns consists of the following steps shown in Figure 1:

1. Given a subgroup pattern s, we first extract the characteristic factors given by a set of selectors $F \subseteq \Omega_E$.
2. Next, we score the obtained characteristic subgroup factors F: For each selector $e \in F$ we obtain its respective (confirmation) strength with respect to the target concept. The assigned scores are then mapped to weights denoting the importance of the respective factors.
3. After that, we apply a case-based retrieval method. Concerning the cases contained in the subgroup we retrieve either typical or extreme cases with a high coverage of the characteristic factors F – as exemplifying cases for the subgroup pattern. In the retrieval method the factors can be weighted according to their relative importance, i.e., according to the assigned weights, depending on the requirements of the user.
4. Finally, we merge the retrieved cases in a virtual prototypical pattern case and present this case to the user to facilitate an easier interpretation.

This process is shown in Figure 1. It is incremental and can include user feedback: The user can optionally inspect, select and refine the set of characteristic factors that are considered in the scoring and the retrieval step. Furthermore, the user can also optionally inspect a preview of the retrieved cases before the prototypical pattern case is generated, and can refine or extend this set as well, if needed.

A prototypical pattern case contains both the set of the (scored) characteristic factors, the set of the relevant (retrieved) *subcases*, and other selected factors obtained from the set of subcases. The prototypical pattern case representation serves several purposes:

- The user usually first considers the different factors (with assigned confirmation strengths) of the prototypical pattern case. The case contains the most important factors that characterize the subgroup pattern reflected by the collection of subcases. In that sense, the prototypical pattern case can be regarded as an extended representative case: It can either contain a summary of the typical problem setting of the subgroup, or a range of the extreme settings of the subgroup pattern.
- Furthermore, the set of the typical or extreme cases of the subgroup can be inspected in detail by the user: The prototypical pattern case also contains a mapping from each contained subcase to the set of the most similar subcases in order to identify clusters representing related situations in a specific context.
- Each factor contained in the prototypical pattern case is also linked to the originating (real) cases contained in the case base. Then, the different real world situations in which the factors occurs can be inspected by the user. Furthermore, these links provide the opportunity to locate other relevant meta-information.

In the following sections, we first show how we score and rank the characteristic factors: For each factor we measure the individual importance for confirming the target concept in the subgroup. After that, we describe the case-based techniques for characterizing and exemplifying subgroup patterns in terms of cases, utilizing methods from case-based reasoning. Finally, we describe how to generate prototypical pattern cases.

3.1 Scoring Subgroup Factors

After the set of characteristic factors has been determined, it can already be used for characterizing a subgroup pattern. However, by analyzing these factors further, we can additionally estimate the strength of each supporting factor with respect to the target concept.

In the following we describe a technique for computing confirmation strengths (weights) for the set of characteristic factors F of a given subgroup. To facilitate an easier interpretation by the user, we focus on a restricted set of symbolic categories. We essentially measure the individual strength of a factor $e \in F$ with respect to the evidence it provides for the target concept in the subgroup. It is easy to see that the principal factors will always obtain the strongest confirmation category, while often weaker categories will be assigned to the supporting factors.

For rating the subgroup factors concerning their confirmation strengths, we compare two populations: The true positives contained in the subgroup and the false positives of the total population. In this way we identify how significantly a selector can discriminate between the cases containing the target concept in the subgroup, and all remaining non-target class cases. For example, in the medical domain we would like to identify factors that are characteristic for a subpopulation of all the patients with a certain disease compared to all the healthy patients.

For scoring the characteristic subgroup factors we rely on an adaptation of a method presented in [Atzmueller *et al.*, 2006b]: Given a subgroup, a characteristic factor, and the target concept, we construct a 2×2 contingency table similar to the technique for identifying the supporting factors. We then compare the distribution of the factor of the true positives in the subgroup, i.e., the target class cases, to all negative cases. By definition, this association is always significant concerning the characteristic factors. Next, we compute a score $s \in [0; 1]$ according to the strength of the association using the ϕ-coefficient for binary variables (c.f., [Atzmueller *et al.*, 2006b]), utilizing the generated contingency table.

Next, there are two options for utilizing the score: First, we can map the obtained score to a symbolic confirmation category $sc \in \{+, ++, +++\}$ that specifies confirming symbolic categories in ascending order using a suitable conversion table. The symbolic category sc expresses the strength or the relative importance of a given selector e. For each factor (selector) $e \in F$ we construct a scoring selector $e' = (e, sc)$ assigning the respective confirmation category sc. Then, we can present the scored selectors to the user for an intuitive overview of the important factors and their corresponding strength for confirming the target concept of the subgroup. Second, we can utilize the obtained scores for the case-based retrieval method described below: Since the confirmation categories denote the strength of the association between an individual factor and the target concept of the subgroup, we can directly map the individual categories to weights denoting the relative importance of the factors. The weights can then be applied in the retrieval method when estimating the similarity of cases.

3.2 Identifying Exemplary Cases for Subgroup Patterns

As a first step for analyzing a specific subgroup pattern we retrieve a set of exemplary cases of the pattern: In this way, we aim to utilize the implicit experiences contained in the cases of the case base as explaining examples. Given a set of characteristic factors F of the subgroup or a user-selected subset of these, either typical or extreme cases with a high coverage of the set of factors F can be retrieved. By inspecting these sets of cases 'as is' the user is already able to obtain a view on the general 'problem setting' of the subgroup. The next step combines these cases and the factors into a prototypical case as an intuitive alternative form. In the next section we describe how the factors and the cases are merged into a prototypical pattern case as a condensed representation.

For exemplifying a subgroup pattern, a naive solution retrieves all the target class cases contained in the subgroup. However, this approach suffers from two shortcomings: First, the set of cases can be quite large for a comprehensive overview. Furthermore, a subset of F is not accounted for very precisely, i.e., the supporting factors: The target class cases contained in the subgroup are determined by the set of principal factors contained in the subgroup, and the target concept only. In contrast to only considering the subgroup description, the set of supporting factors might cover quite a diverse set of cases, since they are not contained in all of the cases. During the retrieval step, we can take the individual strengths of the factors into account utilizing the learned weights. Additionally, we can also include other background knowledge, e.g., partial similarities between attribute values, if available.

Case Retrieval We aim to retrieve a set of (target-class) cases contained in the subgroup that have a high coverage with the set $F \subseteq \Omega_E$ containing the characteristic factors (selectors). Then, we have two options to characterize the set F: First we can retrieve *typical* cases that are most similar to F while the individual cases can also be very similar to each other. These cases can then be used to exemplify the most common factors contained in F. Second, we can retrieve *extreme* cases, i.e., cases that are very similar to F but not to each other. This set of diverse cases is discriminative and can be used in order to obtain a comprehensive view on the setting of extreme factor combinations concerning the set F.

For the retrieval step we use retrieval techniques adapted from case-based reasoning methods [Aamodt and Plaza, 1994]. Given a query case q, we aim to retrieve the k most similar cases $\{c_1, \ldots, c_k\}, c_i \in CB$. The attribute values contained in the query case are commonly called the *problem description*. We consider a *virtual* query case q and define its problem description as the set of characteristic factors F_i obtained from a given subgroup s_i. Optionally, the user can modify and tune F_i interactively to fit the analysis requirements. For example, a subset F' of the factors F_i can be selected, e.g., the most interesting factors. Furthermore, the analysis can also be extended to the non-target class cases contained in the subgroup. Thus, specific queries can be easily formulated by the user.

For assessing the similarity of a (generated) query case q and a retrieved case c, we can use the well-known *matching features* similarity function $sim(q, c)$ given in Equation 1:

$$sim(q, c) = \frac{|\{e \in F' : \pi_e(q) = \pi_e(c)\}|}{|F'|}, \qquad (1)$$

for which we consider the factors $F' \subseteq F_i$ contained in the query case q; $\pi_e(c)$ returns the value corresponding to selector $e = \big(a, \{v\}\big)$, i.e., v for a (virtual) query case c, and the value of the corresponding attribute a otherwise.

Additionally we can apply a weighted similarity measure given in Equation 2 by taking the learned weights of the factors into account, if these factors should not be equally weighted. Additionally, we can apply partial similarities between attribute values, if these are available:

$$sim(c, c') = \frac{\sum\limits_{e \in F'} w(e) \cdot sim\big(\pi_e(c), \pi_e(c')\big)}{\sum\limits_{e \in F'} w(e)}, \qquad (2)$$

where $w(e)$ is the weight of the factor e. If we do not consider partial similarities between attribute values and the weights of factors, then it is easy to see that the formula simplifies to the standard similarity measure given in Equation 1. If partial similarities are not available, then we can define a default similarity of 1 if the factors are equal, and 0 otherwise.

The diversity of a set of retrieved cases $\mathcal{RC} = \{c_i\}_k$ of size k is computed according to the measure $diversity(\mathcal{RC})$, defined as follows:

$$diversity(\mathcal{RC}) = \frac{\sum\limits_{i=1}^{k-1} \sum\limits_{j=i+1}^{k} \big(1 - sim(c_i, c_j)\big)}{k \cdot \frac{(k-1)}{2}}, \qquad (3)$$

where the similarity of two cases is estimated with respect to the attributes in the constructed query case q, as described above.

To retrieve the set of the most extreme cases with respect to a subgroup pattern we apply techniques that obtain a set of most similar but diverse cases regarding to the query case. There are several methods to retrieve a set of diverse cases as described, e.g., in [McSherry, 2002]. We apply the *Bounded Greedy (BG)* algorithm introduced by Smyth and Mc Clave [Smyth and McClave, 2001]: BG starts with a retrieval set initially containing the most similar case to the query case. In each iteration of the algorithm the case in the set of the $2k$ most similar cases is selected which maximizes both the product of its similarity to the query case and its relative diversity with respect to the cases that have been selected for the retrieval set so far.

The relative diversity $relDiversity(c, RC)$ of a case c with respect to the retrieval set $\mathcal{R}C = \{c_i\}_m$ of size m is defined as

$$relDiversity(c, RC) = \frac{\sum_{i=1}^{m} 1 - sim(c, c_i)}{m} . \quad (4)$$

BG stops if the retrieval set reaches its pre-specified size k. To obtain a smaller number of diverse (extreme) cases, we can optionally select the smallest subset $R' \subseteq R$, for which the coverage between the problem description of a query case q and the union of the problem descriptions contained in R' is maximized.

The retrieved set of typical (or extreme) cases can be seen as a set of explaining examples for the given set of factors characterizing a specific subgroup. Thus, a subgroup can be inspected in a different view by considering specific exemplary cases. By presenting typical or extreme cases the user gets a detailed and intuitive impression about the objects (cases) contained in the subgroup. In the next section, we describe how to merge the retrieved cases into a prototypical pattern case for a convenient presentation to the user.

3.3 Generating Prototypical Pattern Cases

In order to create a representative of the retrieved typical or extreme cases of a subgroup pattern, we construct a *prototypical pattern case*. The prototypical pattern case is created by merging the set of the retrieved subcases or a user-selected subset of these. Basically, we then need to combine the contained attribute values and meta-information of the individual subcases.

A *prototypical pattern case*

$$cp = (\mathcal{V}_{cp}, C_{cp}, \sigma_{cp}, \delta_{cp})$$

consists of a set of subcases $C_{cp} \subseteq CB$ of a given case base CB, a set of attribute values $\mathcal{V}_{cp}$ generated using the subcases, a mapping function from an attribute value of the constructed prototypical to its set of (originating) subcases

$$\sigma_{cp} : \mathcal{V}_{cp} \rightarrow 2^{C_{cp}} ,$$

and a selection function

$$\delta_{cp} : C_{cp} \times \mathbb{N} \rightarrow 2^{C_{cp}} .$$

The selection function δ_{cp} retrieves a set of the most similar l subcases for a specific subcase of the prototypical pattern case cp, for which $l \in \mathbb{N}, l \leq k = |C_{cp}|$.

When combining the parts of the subcases, we can simply merge the contained meta-information $\mathcal{I}_c$ of each subcase c. The set of attribute values $\mathcal{V}_{cp} \subseteq \mathcal{V}_A$ of the prototypical pattern case is basically created by joining the attribute values of the subcases:

$$\mathcal{V}_{cp} = \bigcup_{c \in C_{cp}} \mathcal{V}_c .$$

It is easy to see that we can transfer all the attribute values included in the query case to the prototypical pattern case: These factors are given by the set of characteristic (principal and supporting) factors (or a user-selected subset of these) and should therefore always be contained in the prototypical pattern case. For the remaining attributes not included in the set of characteristic attributes we need to select a discriminative set of *additional attribute values* contained in the set of subcases. However, when combining the problem descriptions, i.e. sets of attribute values, conflicts can arise if two cases contain different values for the same attribute. Therefore, we need to apply a conflict resolution step for competing attribute values for a specific attribute, i.e., if $\mathcal{V}_{cp}$ contains more than one attribute value for an attribute.

The following algorithm implements such a conflict resolution strategy for determining the set of additional attribute values during the merge step:

1. We choose the value contained in the query case if included in one subcase.
2. Otherwise, we either draw a majority vote or we can apply background knowledge, if available:
   (a) Generally, we select the most frequently occurring value v from the set of the respective attribute values V contained in the subcases, i.e.,
   $$v = \arg\max_{v_i}(freq\{v_i \in V\}) .$$
   In the case of ties, we select the value that is associated most positively with the target concept utilizing the technique described in Section 3.1.
   (b) Alternatively, we can apply background knowledge, if available: Utilizing partial similarities between attribute values we can select the value which is most similar to the value included in the query case.
   Additionally, we can utilize *abnormality knowledge* (e.g., [Atzmueller *et al.*, 2005b]) which is quite common in some domains, e.g., in the medical domain. Abnormality knowledge specifies which attribute values represent a normal or an abnormal state of their corresponding attribute, e.g. the value *pain=none* is normal, whereas *pain=high* is abnormal for a certain attribute/symptom. If abnormalities are defined, then we select the value with the highest abnormality. This approach is motivated by the heuristic that often especially the abnormal values are interesting, e.g., in the medical domain. For example, if we consider two patients with two (different) diseases, then it seems to be reasonable that the more severe attribute value (finding) will be selected, e.g. *pain=high* from one diagnosis rather than *pain=none* from another one. This is especially helpful when considering a set of extreme cases characterizing the subgroup, since the abnormal values indicate extreme conditions.

The set of attribute values of a generated prototypical pattern case is then given by the set of principal factors and supporting factors of a given subgroup pattern, and by additional factors contained in a set of exemplifying cases.

We model the mapping function σ_{cp} of a prototypical pattern case cp by creating a link from each attribute value of the case cp to the set of the original subcases containing the value, when merging the set of attribute values.

Both the selection and the mapping function enable a 'drill-down' approach when further analyzing a set of factors or a set of cases: The user can easily inspect a related set of subcases, and can also inspect each originating case for a specific attribute value.

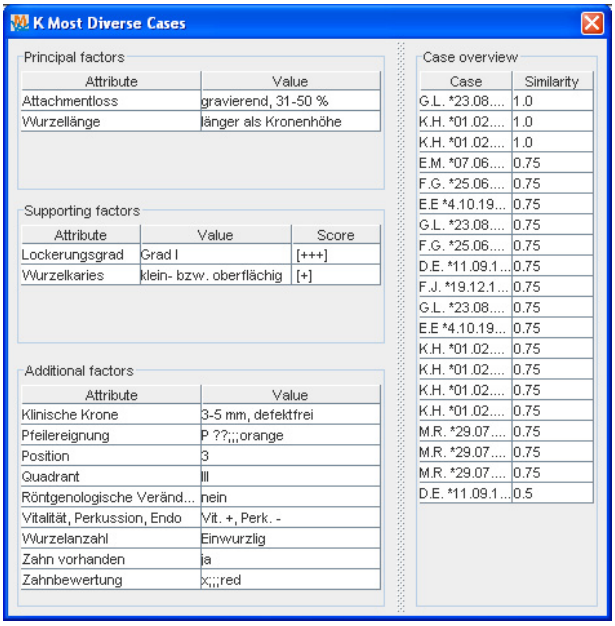

Figure 2: A Prototypical Pattern case for the subgroup *attachmentloss=strong AND root length=longer than crown length* (with respect to the target concept *incorrect tooth extraction*). The left pane contains the principal factors, the supporting factors (*toothlax=minor, root caries=minor*) and their associated scores, and the additional factors of the prototypical pattern case. The right pane shows the retrieved subcases, i.e., in this example the 20 most diverse cases for the given subgroup pattern.

Figure 2 shows an exemplary screenshot of a prototypical pattern case for the domain of dental medicine: The figure depicts the subgroup *attachmentloss=strong AND root length=longer than crown length*, and shows the principal factors, the supporting factors and their strengths, other additional factors, and the set of subcases of the generated prototypical pattern case.

3.4 Discussion

Characterizing subgroup patterns by a set of supporting factors has been proposed by Gamberger et al. [Gamberger *et al.*, 2005; 2003]. The methods for obtaining the supporting factors and for ranking these can be regarded as being related to correlation-based methods for relevance analysis of attributes and attribute values. However, in comparison to such approaches for estimating the importance or the relevance of attribute values (e.g., [Hall, 2000]) and for learning weights of attributes (e.g., [Aha, 1992]) in a case-based reasoning context, the supporting factors focus on descriptive aspects of a subgroup pattern. Thus, the importance of the attributes is estimated with respect to a pattern and a specific target concept, and not concerning the class only: The supporting factors can characterize the subgroup in a different way, orthogonal to the subgroup description.

In contrast to only obtaining the supporting factors (and thus also a subset of the characteristic factors), we further rank these in order to obtain their confirmation strength for the target concept. The obtained confirmation strengths are given by symbolic categories in order to enable an intuitive interpretation for the user. Furthermore, we can directly map these to weights (denoting their relative importance) for the similarity measure used in the case-based retrieval method.

Using prototypical cases has been introduced early in the field of case-based reasoning, e.g., [Bareiss, 1989], and is often applied in medical domains [Schmidt and Gierl, 2001]. In contrast to the existing approaches, we do not just aim at summarizing or describing a set of cases. Instead, we focus on characterizing subgroup patterns: First, we obtain characteristic factors using statistical analysis. Using these we retrieve sets of exemplifying cases. After that, we combine both into a prototypical pattern case, a process for which we can include background knowledge, if available. This prototypical pattern case then provides a comprehensive and condensed alternative representation of a subgroup pattern in the form of a single case.

The different components of a prototypical pattern case $cp = (\mathcal{V}_{cp}, C_{cp}, \sigma_{cp}, \delta_{cp})$ can then be applied in order to fulfill the requirements sketched above in Section 3:

- The set $\mathcal{V}_{cp}$ provides for a condensed form of the characteristic factors and a summary of the remaining factors contained in the subcases. In this way, the generated case can be seen as an alternative view of a subgroup pattern.
- The links between the factors contained in the problem description $\mathcal{V}_{cp}$ of the prototypical pattern case and its set of subcases C_{cp} provide an easy approach for inspecting the important factors in their specific context, i.e., embedded in their originating cases. Furthermore, associated meta-information can be conveniently identified.
- The set C_{cp} and the selection function δ_{cp} facilitate an easy inspection and traversal of the neighborhood of exemplifying cases with respect to the given subgroup pattern. Then, also relevant meta-information can be located quite easily.

So, a prototypical pattern case provides for a concise, easy to interpret, and transparent representation for analyzing, summarizing and characterizing a specific subgroup pattern. Starting with the problem description of a prototypical pattern case, the user can always apply 'drill-down' techniques in order to obtain additional information.

4 Application – Case Studies

The presented approach has already been successfully applied in medical domains. In the following we sketch two case studies: The first case study is given by an application in the domain of sonography utilizing cases from the SONOCONSULT system: Subgroup discovery is applied as a technique for knowledge discovery and for quality control. Then, the discovered subgroup patterns could be conveniently analyzed using the case-based techniques.

The second case study was performed with respect to a knowledge refinement setting applying subgroup discovery techniques in the domain of dental medicine. The goal was to improve a given knowledge-base by analyzing subgroup patterns denoting patterns with a high share of erroneous diagnoses. Then, the knowledge base could be extended by modifying and adding new relations (rules) as needed.

4.1 Characterizing Subgroup Patterns in the Context of Knowledge Discovery

For the first case study, we applied cases acquired using the SONOCONSULT [Huettig *et al.*, 2004] system. SONOCONSULT is a medical documentation and consultation system for sonography which has been developed with the knowledge system D3 [Puppe, 1998].

SONOCONSULT is in routine use in the DRK-hospital in Berlin/Köpenick and in the Würzburg University Hospital. The documented cases contain detailed descriptions of findings of the examination(s), together with the inferred diagnoses, and additional meta-information. The derived diagnoses of a case are usually correct as shown in a medical evaluation, c.f. [Huettig *et al.*, 2004], resulting in a high-quality case base with detailed case descriptions. Currently, the collected SONOCONSULT case base consists of about 11,000 cases. Due to the structured data gathering strategy and the high quality of the case descriptions the system and the collected case base provide excellent opportunities for data analysis and knowledge discovery.

We already utilized parts of the collected case base of SONOCONSULT for knowledge discovery and for data analysis using subgroup mining methods, e.g., [Atzmueller *et al.*, 2005b; 2005c]. The methods were applied in order to discover interesting clinical relations between different organ systems since the inter-organ relations are usually known in the domain of sonography. Furthermore, we applied subgorup mining for quality control with respect to the documentation habits of the sonographic examiners. Then, novel relations between different organ systems could be discovered and documentation profiles for certain examiners could be obtained. Both the relations and the profiles are represented by interesting subgroup patterns.

However, after performing knowledge discovery, the demand for a deeper inspection and characterization of the discovered subgroup patterns in terms of real cases and the further need for identifying related meta-information contained in the cases motivated the development of the presented techniques. Concerning these, the proposed methods for characterization and analysis of subgroup patterns provide powerful opportunities: The medical experts could directly locate interesting contexts, i.e., exemplary cases of specific patients, and typical case descriptions for a specific subgroup pattern. The generated prototypical cases were applied in order to obtain a summary of the typical problem setting of a subgroup pattern, and for subsequently identifying relevant meta-information contained in the characteristic set of cases. Concerning the case-studies the users could easily discover relevant meta-information, e.g., certain examiners and images associated with a given subgroup pattern using the prototypical case; the location of specific images of sonographic situations proved especially interesting for the medical clinicians.

4.2 Analyzing Subgroup Patterns in the Context of Interactive Knowledge Refinement

The second case study concerns the domain of dental medicine were we used subgroup mining for interactive knowledge refinement of a knowledge-based system. The case study was performed in the domain of dental medicine implemented with a consultation and documentation system for dental findings regarding any kind of prosthetic appliance. The system has been developed in cooperation with the department of prosthodontics at the Würzburg University Hospital.

The system aims to decide about a diagnostic plan using the clinical findings: The cases always contain the standard anamnestic findings and additional findings from x-ray examinations, e.g., abnormal x-ray findings (apical, periradicular), grade of tooth lax, endodontic state (root filling, pulp vitality), root quantity, root length, crown length, level of attachment loss, root caries, tooth angulation and elongation/extrusion. For decision support the system derives two distinct diagnosis *EX* and *IN* that either indicate the teeth that could be conserved (IN) or should be extracted (EX).

We successfully applied a method for knowledge-refinement using subgroup mining methods, in order to improve the correctness of the knowledge base that initially was in an earlier state. Therefore, we were able to improve the knowledge base significantly by adding and modifying relations that were identified using a subgroup mining approach, c.f., [Atzmueller *et al.*, 2005a; 2006a]. Subgroup mining was applied for pointing at certain subgroups corresponding to 'hot spots' of the knowledge base, i.e., specific factor combinations for which the error rate of the system increased significantly. These subgroups were then analyzed by the domain specialists in order to perform refinement operators on the knowledge base, e.g., modifying relations or adding new ones.

However, the experiences obtained throughout the earlier parts of the case study motivated the development of further methods for subgroup characterization, introspection and analysis: Often small 'hot spots', i.e., very specific subgroup patterns, needed to be analyzed in detail, either statistically or by viewing the detailed cases.

The presentation of characterizing subgroup factors and a set of exemplifying cases merged to prototypical pattern cases was a key feature for the domain specialist, who performed the analysis. Figure 2 in Section 3.3 shows an example of such a case. The method allowed for a comprehensive overview on the sub-population defined by a small set of exemplary cases. Especially interesting were the summarization and presentation of the characteristic factors by a prototypical pattern case, and the 'drill-down' options into exemplifying cases. Especially the drill-down techniques from factors to sets of cases and for navigating the neighborhood of the a set of retrieved cases proved very helpful during the application. This provided for an easier analysis of the important factor combinations, their contributions and the specific contexts they occurred in.

5 Conclusion

In this paper we have introduced case-based methods for subgroup analysis and characterization. Combining these, we presented an approach that first characterizes a subgroup in terms of its characteristic factors, ranks these, retrieves corresponding typical or extreme cases and finally combines both into a prototypical pattern case. Using this representation, the user can get a comprehensive overview of the problem setting of the subgroup pattern. Furthermore, using 'drill-down' operations on the set of cases, further interesting meta-information contained in the characteristic (real) cases can be identified. We can apply several types of background knowledge during the merge step, depending on the requirements of the user.

In the future, we plan to investigate further techniques for subgroup characterization and summarization, e.g., based on clustering techniques, and also regarding other condensed forms of sets of subgroups.

Acknowledgements

We want to thank Achim Hemsing and Prof. Ernst-Jürgen Richter from the department of prosthodontics at the Würzburg University Hospital, Prof. Hans-Peter Buscher from the department of internal medicine at the DRK-Klinik Berlin/Köpenick, and Hardi Lührs from the department of sonography at the Würzburg University Hospital for their medical expertise and analysis while performing the case studies of this research project.

References

[Aamodt and Plaza, 1994] Agnar Aamodt and Enric Plaza. Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Communications*, 7(1):39 – 59, 1994.

[Aha, 1992] David W. Aha. Tolerating Noisy, Irrelevant and Novel Attributes in Instance-Based Learning Algorithms. *Intl. Journal of Man-Machine Studies*, 36(2):267–287, 1992.

[Atzmueller *et al.*, 2005a] Martin Atzmueller, Joachim Baumeister, Achim Hemsing, Ernst-Jürgen Richter, and Frank Puppe. Subgroup Mining for Interactive Knowledge Refinement. In *Proc. 10th Conference on Artificial Intelligence in Medicine (AIME 05)*, LNAI 3581, pages 453–462, Berlin, 2005. Springer.

[Atzmueller *et al.*, 2005b] Martin Atzmueller, Frank Puppe, and Hans-Peter Buscher. Exploiting Background Knowledge for Knowledge-Intensive Subgroup Discovery. In *Proc. 19th Intl. Joint Conference on Artificial Intelligence (IJCAI-05)*, pages 647–652, Edinburgh, Scotland, 2005.

[Atzmueller *et al.*, 2005c] Martin Atzmueller, Frank Puppe, and Hans-Peter Buscher. Profiling Examiners using Intelligent Subgroup Mining. In *Proc. 10th Intl. Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-2005)*, pages 46–51, Aberdeen, Scotland, 2005.

[Atzmueller *et al.*, 2006a] Martin Atzmueller, Joachim Baumeister, and Frank Puppe. Introspective Subgroup Analysis for Interactive Knowledge Refinement. In Geoff Sutcliffe and Randy Goebel, editors, *Proc. 19th Intl. Florida Artificial Intelligence Research Society Conference 2006 (FLAIRS-2006)*, pages 402–407. AAAI Press, 2006.

[Atzmueller *et al.*, 2006b] Martin Atzmueller, Joachim Baumeister, and Frank Puppe. Semi-Automatic Learning of Simple Diagnostic Scores utilizing Complexity Measures. *Artificial Intelligence in Medicine. Special Issue on Intelligent Data Analysis in Medicine*, 37(1):19–30, 2006.

[Bareiss, 1989] Ray Bareiss. *Exemplar-Based Knowledge Acquisition: A Unified Approach to Concept Representation, Classification, and Learning*. Academic Press Professional, Inc., San Diego, CA, USA, 1989.

[Bartsch-Spörl *et al.*, 1999] Brigitte Bartsch-Spörl, Mario Lenz, and André Hübner. Case-Based Reasoning: Survey and Future Directions. In *XPS-99: Knowledge-Based Systems - Survey and Future Directions, Proc. 5th Biannual German Conference on Knowledge-Based Systems*, pages 67–89, 1999.

[Bergmann, 2002] Ralph Bergmann. *Experience Management: Foundations, Development Methodology, and Internet-Based Applications*. Springer, Berlin, 2002.

[Gamberger *et al.*, 2003] Dragan Gamberger, Nada Lavrac, and Goran Krstacic. Active Subgroup Mining: A Case Study in Coronary Heart Disease Risk Group Detection. *Artificial Intelligence in Medicine*, 28:27–57, 2003.

[Gamberger *et al.*, 2005] Dragan Gamberger, Antonija Krstacic, Goran Krstacic, Nada Lavrac, and Michele Sebag. Data Analysis Based on Subgroup Discovery: Experiments in Brain Ischaemia Domain. In *Proc. 10th Intl. Workshop on Intelligent Data Analysis in Medicine and Pharmacology (IDAMAP-2005)*, pages 52–56, Aberdeen, Scotland, 2005.

[Hall, 2000] Mark A. Hall. Correlation-based Feature Selection for Discrete and Numeric Class Machine Learning. In *Proc. 17th Intl. Conf. on Machine Learning*, pages 359–366. Morgan Kaufmann, San Francisco, CA, 2000.

[Huettig *et al.*, 2004] Matthias Huettig, Georg Buscher, Thomas Menzel, Wolfgang Scheppach, Frank Puppe, and Hans-Peter Buscher. A Diagnostic Expert System for Structured Reports, Quality Assessment, and Training of Residents in Sonography. *Medizinische Klinik*, 99(3):117–122, 2004.

[Klösgen, 1996] Willi Klösgen. Explora: A Multipattern and Multistrategy Discovery Assistant. In Usama M. Fayyad, Gregory Piatetsky-Shapiro, Padraic Smyth, and Ramasamy Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, pages 249–271. AAAI Press, 1996.

[Lavrac *et al.*, 2002] Nada Lavrac, Dragan Gamberger, and Peter Flach. Subgroup Discovery for Actionable Knowledge Generation: Shortcomings of Classification Rule Learning and the Lessons Learned. In Nada Lavrac, Hiroshi Motoda, and Tom Fawcett, editors, *Proc. ICML 2002 workshop on Data Mining: Lessons Learned*, July 2002.

[McSherry, 2002] David McSherry. Diversity-Conscious Retrieval. In *Proc. 6th European Conference on Advances in Case-Based Reasoning*, pages 219–233, Berlin, 2002. Springer.

[Puppe, 1998] Frank Puppe. Knowledge Reuse among Diagnostic Problem-Solving Methods in the Shell-Kit D3. *Intl. Journal of Human-Computer Studies*, 49:627–649, 1998.

[Schmidt and Gierl, 2001] Rainer Schmidt and Lothar Gierl. Case-based Reasoning for Antibiotics Therapy Advice: An Investigation of Retrieval Algorithms and Prototypes. *Artificial Intelligence in Medicine*, 23(2):171–186, 2001.

[Smyth and McClave, 2001] Barry Smyth and Paul McClave. Similarity vs. Diversity. In *Proc. 4th Intl. Conference on Case-Based Reasoning (ICCBR 01)*, pages 347–361, Berlin, 2001. Springer.

[Wrobel, 1997] Stefan Wrobel. An Algorithm for Multi-Relational Discovery of Subgroups. In Jan Komorowski and Jan Zytkow, editors, *Proc. 1st European Symposium on Principles of Data Mining and Knowledge Discovery (PKDD-97)*, pages 78–87, Berlin, 1997. Springer.